

# Ensemble Sparse Models for Image Analysis

Karthikeyan Natesan Ramamurthy, Jayaraman J. Thiagarajan, Prasanna Sattigeri, and Andreas Spanias

**Abstract**—Sparse representations with learned dictionaries have been successful in several image analysis applications. In this paper, we propose and analyze the framework of ensemble sparse models, and demonstrate their utility in image restoration and unsupervised clustering. The proposed ensemble model approximates the data as a linear combination of approximations from multiple *weak* sparse models. Theoretical analysis of the ensemble model reveals that even in the worst-case, the ensemble can perform better than any of its constituent individual models. The dictionaries corresponding to the individual sparse models are obtained using either random example selection or boosted approaches. Boosted approaches learn one dictionary per round such that the dictionary learned in a particular round is optimized for the training examples having high reconstruction error in the previous round. Results with compressed recovery show that the ensemble representations lead to a better performance compared to using a single dictionary obtained with the conventional alternating minimization approach. The proposed ensemble models are also used for single image superresolution, and we show that they perform comparably to the recent approaches. In unsupervised clustering, experiments show that the proposed model performs better than baseline approaches in several standard datasets.

**Index Terms**—Sparse coding, dictionary learning, ensemble models, image recovery, clustering.

## I. INTRODUCTION

Natural signals and images reveal statistics that allow them to be efficiently represented using a sparse linear combination of elementary patterns [1]. The local regions of natural images, referred to as patches, can be represented using a sparse linear combination of columns from a dictionary matrix. Given a data sample  $\mathbf{x} \in \mathbb{R}^M$ , and a dictionary matrix  $\mathbf{D} \in \mathbb{R}^{M \times K}$ , the data approximated using the linear generative model as  $\mathbf{D}\mathbf{a}$ , where  $\mathbf{a} \in \mathbb{R}^K$  is the sparse coefficient vector. This generative model that incorporates sparsity constraints in the coefficient vector, will be referred to as the *sparse model*, in this paper. The dictionary can be either pre-defined or learned from the training examples themselves. Learning the dictionary will be alternatively referred to as learning the sparse model. Learned dictionaries have been shown to provide improved performance for restoring degraded data in applications such as denoising, inpainting, deblurring, superresolution, and compressive sensing [2], [3], and also in machine learning applications such as classification and clustering [4]–[6].

### A. Sparse Coding and Dictionary Learning

Using the linear generative model, the sparse code of a data sample  $\mathbf{x}$  can be obtained by optimizing,

$$h(\mathbf{x}, \mathbf{D}) = \min_{\mathbf{a}} \|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2^2 + \lambda \|\mathbf{a}\|_1. \quad (1)$$

Here  $\|\mathbf{a}\|_1$  is the  $\ell_1$  penalty that promotes sparsity of the coefficients, and the equivalence of (1) to  $\ell_0$  minimization has been discussed in [7] under some strong conditions on the dictionary  $\mathbf{D}$ . Some methods to obtain sparse representations used include the Matching Pursuit (MP) [8], Orthogonal Matching Pursuit (OMP) [9], Order-Recursive Matching Pursuit [10], Basis Pursuit (BP) [11], FOCUSS [12] and iterated shrinkage algorithms [13], [14].

In several image processing and sample machine learning applications, it is advantageous to learn a dictionary, such that the set of training examples obtained from a probability space have a small approximation error with sparse coding. This problem can be expressed as minimizing the objective [15]

$$g(\mathbf{D}) = \mathbf{E}_{\mathbf{x}}[h(\mathbf{x}, \mathbf{D})], \quad (2)$$

where the columns of  $\mathbf{D}$ , referred to as dictionary atoms, are constrained to have unit  $\ell_2$  norm, i.e.,  $\|\mathbf{d}_j\|_2 \leq 1, \forall j$ . If the distribution in the probability space is unknown and we only have  $T$  training examples  $\{\mathbf{x}_i\}_{i=1}^T$ , each with probability mass  $p(\mathbf{x}_i)$ , (2) can be modified as the empirical cost function,

$$\hat{g}(\mathbf{D}) = \sum_{i=1}^T h(\mathbf{x}_i, \mathbf{D})p(\mathbf{x}_i). \quad (3)$$

Typically dictionary learning algorithms solve for the sparse codes [16], [17] using (1), and obtain the dictionary by minimizing  $\hat{g}(\mathbf{D})$ , repeating the steps until convergence. We refer to this baseline algorithm as *Alt-Opt*. Since this is an alternating minimization process, it is important to provide a good initial dictionary and this is performed by setting the atoms to normalized cluster centers of the data [18]. Instead of learning dictionaries using sophisticated learning algorithms, it is possible to use the training examples themselves as the dictionary. Since the number of examples  $T$  is usually much larger than the number of dictionary atoms  $K$ , it is much more computationally intensive to obtain sparse representations with examples. Nevertheless, both learned and example-based dictionaries have found applications in inverse problems [2], [19], [20] and also in machine learning applications such as clustering and classification [4], [5], [21]–[29].

### B. Ensemble Sparse Models

In this paper, we propose and explore the framework of ensemble sparse models, where we assume that data can be represented using a linear combination of  $L$  different sparse approximations, instead of being represented using an approximation obtained from a single sparse model. The approximation to  $\mathbf{x}$  can be obtained by optimizing

$$\min_{\{\beta_l\}_{l=1}^L} \left\| \mathbf{x} - \sum_{l=1}^L \beta_l \mathbf{D}_l \mathbf{a}_l \right\|_2^2. \quad (4)$$

The authors are with the SenSIP Center, School of ECEE, Arizona State University, USA. 85287-5706

Email: {knatesan, jjayaram, psattiger, spanias}@asu.edu

Here each coefficient vector  $\mathbf{a}_l$  is assumed to be sparse, and is obtained by solving for the optimization (1) with  $\mathbf{D}_l$  as the dictionary. The weights  $\{\beta_l\}_{l=1}^L$  control the contribution of each base model to the ensemble.

Since the ensemble combines the contributions of multiple models, it is sufficient that the dictionary for model is obtained using a “weak” training procedure. We propose to learn these *weak dictionaries*  $\{\mathbf{D}_l\}_{l=1}^L$  sequentially, using a greedy forward selection procedure, such that training examples that incurred a high approximation error with the dictionary  $\mathbf{D}_l$  are given more importance while learning  $\mathbf{D}_{l+1}$ . Furthermore, we also propose an ensemble model where each individual dictionary is designed as a random subset of training samples. The formulations described in this paper belong to the category of boosting [30] and random selection algorithms [31] in machine learning. In supervised learning, boosting is used to improve the accuracy of learning algorithms, using multiple weak hypotheses instead of a single strong hypothesis. The proposed ensemble sparse models are geared towards two image analysis problems, the inverse problem of restoring degraded images, and the problem of unsupervised clustering. Note that, boosted ensemble models have been used with the bag-of-words approach for updating codebooks in classification [32] and medical image retrieval [33]. However, it has not been used so far in sparsity based image restoration problems or unsupervised clustering. Also when compared to [34], where the authors propose to obtain multiple randomized sparse representations from a single dictionary, in our approach, we propose to learn an ensemble of dictionaries and obtain a single representation from each of them. Typical ensemble methods for regression [35] modify the samples in each round of leveraging, whereas in our case the same training set is used for each round.

### C. Contributions

In this work, we propose the framework of ensemble sparse models and perform a theoretical analysis that relates their performance when compared to its constituent base sparse models. We show that, even in the worst case, an ensemble will perform at least as well as its best constituent sparse model. Experimental demonstrations that support this theory are also provided. We propose two approaches for learning the ensemble: (a) using a random selection and averaging (*RandExAv*) approach, where each dictionary is chosen as a random subset of the training examples, and (b) using a boosted approach to learn dictionaries sequentially by modifying the probability masses of the training examples in each round of learning. In the boosted approach, two methods to learn the weak dictionaries for the individual sparse models, one that performs example selection using the probability distribution on the training set (*BoostEx*), and the other that uses a weighted K-means approach (*BoostKM*), are provided. For all cases of ensemble learning, we also provide methods to obtain the ensemble weights,  $\{\beta_l\}_{l=1}^L$ , from the training examples. Demonstrations that show the convergence of ensemble learning, with the increase in the number of constituent sparse models are provided. Experiments also show that the proposed

ensemble approaches perform better than their best constituent sparse models, as predicted by theory.

In order to demonstrate the effectiveness of the proposed ensemble models, we explore its application to image recovery and clustering. The image recovery problems that we consider here are compressive sensing using random projections and single image superresolution. When boosted ensemble models are learned for image recovery problems, the form of degradation operator specific to the application is also considered, thereby optimizing the ensemble for the application. For compressive recovery, we compare the performance of the proposed Random Example Averaging (*RandExAv*), Boosted Example (*BoostEx*), and Boosted K-Means (*BoostKM*) approaches to the single sparse model, whose dictionary is obtained using the *Alt-Opt* approach. It is shown that the ensemble methods perform consistently better than a single sparse model at different number of measurements. Note that, the base sparse model for example-based approaches is designed as a random subset of examples, and hence it requires minimal training. Furthermore, in image superresolution, the performance of the proposed ensemble learning approaches is comparable to the recent sparse representation methods [36], [20].

Furthermore, we explore the use of the proposed approaches in unsupervised clustering. When the data are clustered along unions of subspaces, an  $\ell_1$  graph [29] can be obtained by representing each data sample  $\mathbf{x}_i$  as a sparse linear combination of the rest of the samples in the set. Another approach proposed in [5] computes the sparse coding-based graph using codes obtained with a learned dictionary. We propose to use ensemble methods to compute sparse codes for each data sample, and perform spectral clustering using graphs obtained from them. Results with several standard datasets show that high clustering performance is obtained using the proposed approach when compared to  $\ell_1$  graph-based clustering.

## II. ANALYSIS OF ENSEMBLE MODELS

We will begin by motivating the need for an ensemble model in place of a single sparse model, and then proceed to derive some theoretical guarantees on the ensemble model. Some demonstrations on the performance of ensemble models will also be provided.

### A. Need for the Ensemble Model

In several scenarios, a single sparse model may be insufficient for representing the data, and using an ensemble model instead may result in a good performance. The need for ensemble models in supervised learning have been well-studied [37]. We will argue that the same set of reasons apply to the case of ensemble sparse models also. The first reason is statistical, whereby several sparse models may have a similar training error when learned from a limited number of training samples. However, the performance of each of these models with test data can be poor. By averaging representations obtained from an ensemble, we may obtain an approximation closer to the true test data. The second reason is computational, which can occur with the case of large training sets also.

The inherent issue in this case is that sparse modeling is a problem with locally optimal solution. Therefore, we may never be able to reach the global optimal solution with a single model and hence using an ensemble model may result in a lesser representation error. Note that this case is quite common in dictionary learning, since many dictionary learning algorithms only seek a local optimal solution. The third reason for using an ensemble model is representational, wherein the hypothesis space assumed cannot represent the test data sample. In the case of sparse models, this corresponds to the case where the dictionary cannot provide a high-fidelity sparse approximation for a novel test data sample. This also happens in the case where the test observation is a corrupted version of the underlying test data, and there is ambiguity in obtaining a sparse representation. In this case also, it may be necessary to combine multiple sparse models to improve the estimate of the test data.

In order to simplify notation in the following analysis, let us denote the  $l^{\text{th}}$  approximation in the ensemble model as  $\mathbf{c}_l = \mathbf{D}_l \mathbf{a}_l$ . The individual approximations are stacked in the matrix  $\mathbf{C} \in \mathbb{R}^{M \times L}$ , where  $\mathbf{C} = [\mathbf{c}_1 \dots \mathbf{c}_L]$  and the weight vector is denoted as  $\boldsymbol{\beta} = [\beta_1 \dots \beta_L]^T$ . The individual residuals are denoted as  $\mathbf{r}_l = \mathbf{x} - \mathbf{c}_l$ , for  $i = 1, \dots, L$ , and the total residual of the approximation is given as

$$\mathbf{r} = \mathbf{x} - \mathbf{C}\boldsymbol{\beta}. \quad (5)$$

We characterize the behavior of the ensemble sparse model by considering four different cases for the weights  $\{\beta_l\}_{l=1}^L$ .

1) *Unconstrained Weights*: In this case, the ensemble weights  $\{\beta_l\}_{l=1}^L$  are assumed to be unconstrained and computed using the unconstrained least squares approximation

$$\min_{\boldsymbol{\beta}} \|\mathbf{x} - \mathbf{C}\boldsymbol{\beta}\|_2^2 \quad (6)$$

When the data  $\mathbf{x}$  lies in the span of  $\mathbf{C}$ , the residual will be zero, i.e.,  $\mathbf{r} = 0$ . The residual that has minimum energy in the  $L$  approximations is denoted as  $\mathbf{r}_{\min}$ . This residual can be obtained by setting the corresponding weight in the vector  $\boldsymbol{\beta}$  to be 1, whereas (6) computes  $\boldsymbol{\beta}$  that achieves the best possible residual  $\mathbf{r}$  for the total approximation. Clearly this implies

$$\|\mathbf{r}\|_2 \leq \|\mathbf{r}_{\min}\|_2. \quad (7)$$

Therefore, at worst, the total approximation will be as good as the best individual approximation.

2)  $\beta_l \geq 0$ : The ensemble weights  $\{\beta_l\}_{l=1}^L$  are assumed to be non-negative in this case. The least squares approximation (4), with the constraint  $\boldsymbol{\beta} \geq 0$  will now result in a zero residual if the data  $\mathbf{x}$  lies in the simplicial cone generated by the columns of  $\mathbf{C}$ . The simplicial cone is defined as the set  $\{\mathbf{b} : \mathbf{b} = \sum_{l=1}^L \mathbf{c}_l \beta_l\}$ . Otherwise, the bound on the total residual given by (7) holds in this case, since  $\mathbf{r}_{\min}$  can be obtained by setting the appropriate weight in  $\boldsymbol{\beta}$  to 1 in (5), and the rest to 0 in this case also.

3)  $\sum_{l=1}^L \beta_l = 1$ : When the ensemble weights are constrained to sum to 1, the total residual can be expressed as

$$\mathbf{r} = \sum_{l=1}^L \beta_l \mathbf{r}_l. \quad (8)$$

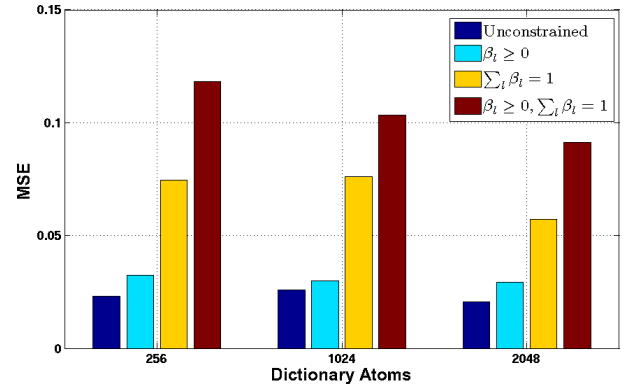


Fig. 1. Performance of the “oracle” ensemble models for various constraints on weights and different dictionary sizes in the base models.

This can be easily obtained by replacing  $\mathbf{x}$  as  $\sum_{l=1}^L \beta_l \mathbf{x}$  in (5). Denoting the residual matrix  $\mathbf{R} = [\mathbf{r}_1 \dots \mathbf{r}_L]$ , the optimization (4) to compute the weights can also be posed as  $\min_{\boldsymbol{\beta}} \|\mathbf{R}\boldsymbol{\beta}\|_2$ . Incorporating the constraint  $\sum_{l=1}^L \beta_l = 1$ , it can be seen that the final approximation  $\mathbf{C}\boldsymbol{\beta}$  lies in the affine hull generated by the columns of  $\mathbf{C}$ , and the final residual,  $\mathbf{R}\boldsymbol{\beta}$ , lies in the affine hull generated by the columns of  $\mathbf{R}$ . Clearly the final residual will be zero, only if the data  $\mathbf{x}$  lies in the affine hull of  $\mathbf{C}$ , or equivalently the zero vector lies in the affine hull of  $\mathbf{R}$ . When this does not hold, the worst case bound on  $\mathbf{r}$  given by (7) holds in this case as well.

4)  $\beta_l \geq 0, \sum_{l=1}^L \beta_l = 1$ : Similar to the previous case, the total residual can be expressed as (8). As a result, the final representation  $\mathbf{C}\boldsymbol{\beta}$  lies in the convex hull generated by the columns of  $\mathbf{C}$ , and the final residual,  $\mathbf{R}\boldsymbol{\beta}$ , lies in the convex hull generated by the columns of  $\mathbf{R}$ . Furthermore, the final residual will be zero only if the zero vector lies in the convex hull of  $\mathbf{R}$ . Clearly, the worst case bound on  $\mathbf{r}$  given by (7) holds in this case.

Although the worst case bounds for all the four cases are the same, the constraint spaces for the cases might provide us an idea about their relative performances with real data. The first case is unconstrained and it should result in the least error. The second case constrains that the solution should lie in the simplicial cone spanned by the columns of  $\mathbf{C}$ , and this should lead to higher residual energy than Case 1. Case 3 constrains the solution to lie in an affine hull, which is of  $L-1$  dimensions compared to simplicial cone in  $L$  dimensions, so it could lead to a higher error compared to Case 2. Case 4 is the subset of constraint spaces for Cases 1 to 3 and hence it will lead to the highest residual error.

## B. Demonstration of Ensemble Representations

In order to demonstrate the performance of ensemble representations with real data, we obtain a random set of 100,000 patches, each of size  $8 \times 8$ , from a set of natural images. The training images were obtained from the superresolution toolbox published by Yang *et al.* [38], and consist of a wide variety of patterns and textures. We will refer to this set of training images simply as the *training image set* throughout this paper. The chosen patches are then processed to remove

the mean, followed by the removal of low-variance patches. Since image recovery is the important application of the proposed models, considering high-variance patches alone is beneficial. Each dictionary in the ensemble  $\mathbf{D}_l$  is obtained as a random set of  $K$  vectorized, and normalized patches. We fix the number of models in the ensemble as  $L = 20$ . The test data is a random set of 1000 grayscale patches obtained from the Berkeley segmentation dataset [39]. For each test sample, we compute the set of  $L$  approximations using the sparse model given in (1), with  $\lambda = 0.2$ . The individual approximations are combined into an ensemble, under the four conditions on the weights,  $\{\beta_l\}$ , described above. The optimal weights are computed and the mean squared norm of the residuals for all the test samples are compared in Figure 1, for the dictionary sizes  $K = \{256, 1025, 2048\}$ . We observe that the performance of the ensembles generally improve as the size of the dictionaries used in the base models increase. The variation in performance across all the four cases of weights follows our reasoning in the previous section. We refer to these as “oracle” ensemble models, since the weights are optimally computed with perfect knowledge of all the individual approximations and the actual data. In reality, the weights will be precomputed from the training data.

### III. PROPOSED ENSEMBLE SPARSE REPRESENTATION ALGORITHMS

The ensemble model proposed in (4) results in a good approximation for any known data. However, in order to use ensemble models in analysis and recovery of images, that are possibly corrupted or degraded, both the weights  $\{\beta_l\}_{l=1}^L$  and the dictionaries  $\{\mathbf{D}_l\}_{l=1}^L$  must be inferred from uncorrupted training data. The set of weights is fixed to be common for all test observations instead of computing a new set of weights for each observation. Let us denote the set of training samples as  $\mathbf{X} = [\mathbf{x}_1 \ \mathbf{x}_2 \ \dots \ \mathbf{x}_T]$ , and the set of coefficients in base model  $l$  as  $\mathbf{A}_l = [\mathbf{a}_{l,1} \ \mathbf{a}_{l,2} \ \dots \ \mathbf{a}_{l,T}]$ , where  $\mathbf{a}_{l,i}$  is the coefficient vector of the  $i^{\text{th}}$  sample for base model  $l$ . In the proposed ensemble learning procedures, we consider both simple averaging and boosting approaches.

#### A. Random Example Averaging Approach

The first approach chooses  $L$  random subsets of  $K$  samples from the training data itself and normalizes them to form the dictionaries  $\{\mathbf{D}_l\}_{l=1}^L$ . The weights  $\{\beta_l\}$  are chosen to be equal for all base models as  $1/L$ . Note that the selection of dictionaries follows the same procedure as given in the previous demonstration (Section II-B). We refer to this ensemble approach as *Random Example Averaging (RandExAv)*.

#### B. Boosting Approaches

The next two approaches use boosting and obtain the dictionaries and weights sequentially, such that the training examples that resulted in a poor performance with the  $l-1^{\text{th}}$  base model are given more importance when learning the  $l^{\text{th}}$  base model. We use a greedy forward selection procedure for obtaining the dictionaries and the weights. In each round  $l$ , the

model is augmented with one dictionary  $\mathbf{D}_l$ , and the weight  $\alpha_l$  corresponding to the dictionary is obtained. The cumulative representation for round  $l$  is given by

$$\mathbf{X}_l = (1 - \alpha_l)\mathbf{X}_{l-1} + \alpha_l\mathbf{D}_l\mathbf{A}_l. \quad (9)$$

Note that the weights of the greedy forward selection algorithm,  $\alpha_l$ , and the weights of the ensemble model,  $\beta_l$ , are related as

$$\beta_l = \alpha_l \prod_{t=l+1}^L (1 - \alpha_t). \quad (10)$$

From (9), it can be seen that  $\mathbf{X}_l$  lies in the affine hull of  $\mathbf{X}_{l-1}$  and  $\mathbf{D}_l\mathbf{A}_l$ . Furthermore, from the relationship between the weights  $\{\alpha_l\}$  and  $\{\beta_l\}$  given in (10), it is clear that  $\sum_{l=1}^L \beta_l = 1$  and hence the ensemble model uses the constraints given in Case 3. Only the Cases 3 and 4 lead to an efficient greedy forward selection approach for the ensemble model in (4), and we use Case 3 since it leads to a better approximation performance (Figure 1).

In boosted ensemble learning, the importance of the training samples in a particular round is controlled by modifying their probability masses. Each round consists of (a) learning a dictionary  $\mathbf{D}_l$  corresponding to the round, (b) computing the approximation for the current round  $l$ , (c) estimating the weight  $\alpha_l$ , (d) computing the residual energy for the training samples, and (e) updating the probability masses of the training samples for the next round. Since the goal of ensemble approaches is to have only *weak* individual models,  $\mathbf{D}_l$  is obtained using naive dictionary learning procedures as described later in this section. The dictionaries for the first round are obtained by fixing uniform probability masses for each training example in the first round, (i.e.),  $p_1(\mathbf{x}_i) = 1/T$  for  $i = \{1, 2, \dots, T\}$ . Assuming that  $\mathbf{D}_l$  is known, the approximation for the current round is computed by coding the training samples  $\mathbf{X}$  with the dictionary using (1). The weight  $\alpha_l$  is computed such that the error between the training samples and the cumulative approximation  $\mathbf{X}_l$  is minimized. Using (9), this optimization can be expressed as

$$\min_{\alpha_l} \|\mathbf{X}_l - [(1 - \alpha_l)\mathbf{X}_{l-1} + \alpha_l\mathbf{D}_l\mathbf{A}_l]\|_F^2, \quad (11)$$

and can be solved in closed form with the optimal value given as,

$$\alpha_l = \frac{\text{Tr}[(\mathbf{X} - \mathbf{X}_{l-1})^T(\mathbf{D}_l\mathbf{A}_l - \mathbf{X}_{l-1})]}{\|\mathbf{D}_l\mathbf{A}_l - \mathbf{X}_{l-1}\|_F^2}, \quad (12)$$

where  $\text{Tr}$  denotes the trace of the matrix. The residual matrix for all the training samples in round  $l$  is given by  $\mathbf{R}_l = \mathbf{X} - \mathbf{D}_l\mathbf{A}_l$ . The energy of the residual for the  $i^{\text{th}}$  training sample is given as  $e_l(i) = \|\mathbf{r}_{l,i}\|_2^2$ . If the dictionary in round  $l$  provides a large approximation error for sample  $i$ , then that sample will be given more importance in round  $l+1$ . This will ensure that the residual error for sample  $i$  in round  $l+1$  will be small. The simple scheme of updating the probability masses as  $p_{l+1}(\mathbf{x}_i) = e_l(i)$ , upweights the badly represented samples and downweights the well-represented ones for the next round.

Given a training set  $\{\mathbf{x}_i\}_{i=1}^L$ , and its probability masses  $\{p_l(\mathbf{x}_i)\}_{i=1}^L$ , we will propose two simple approaches for learning the dictionaries corresponding to the individual sparse models.

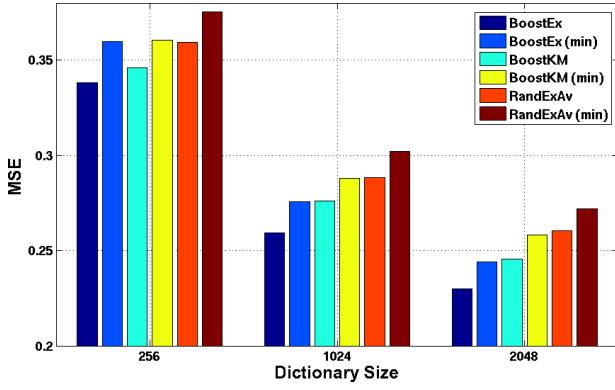


Fig. 2. Comparison of the performance of the proposed ensemble learning approaches for various dictionary sizes.

1) *BoostKM*: When the sparse code for each training example is constrained to take one only one non-zero coefficient of value 1, and the norms of the dictionary atoms are unconstrained, the dictionary learning problem (3) can be shown to reduce to K-Means clustering. Hence, computing a set of K-Means cluster centers and normalizing them to unit  $\ell_2$  norm constitutes a reasonable weak dictionary. However, since the distribution on the data could be non-uniform in our case, we need to alter the clustering scheme to incorporate this. Denoting the cluster centers to be  $\{\mu_k\}_{k=1}^K$ , the cluster membership sets to be  $\{\mathcal{M}_k\}_{k=1}^K$ , the weighted K-Means objective is denoted as

$$\min_{\{\mu_k\}_{k=1}^K, \{\mathcal{M}_k\}_{k=1}^K} \sum_{k=1}^K \sum_{i \in \mathcal{M}_k} p(\mathbf{x}_i) \|\mathbf{x}_i - \mu_k\|_2^2. \quad (13)$$

The weighted K-Means procedure is implemented by modifying the scalable K-Means++ algorithm, also referred to as the K-Means|| (K-Means Parallel) algorithm [40]. The K-Means|| algorithm is an improvement over the K-Means++ algorithm [41] that provides a method for careful initialization leading to improved speed and accuracy in clustering. The advantage with the K-Means|| algorithm is that the initialization procedure is scalable to a large number of samples. In fact, it has been shown in [40] that just the initialization procedure in K-Means|| results in a significant reduction in the clustering cost. Since we are interested in learning only a weak dictionary, we will use the normalized cluster centers obtained after initialization as our dictionary. The K-Means++ algorithm selects initial cluster centers sequentially such that they are relatively spread out. For initializing  $K$  cluster centers, the algorithm creates a distribution on the data samples and picks a cluster center by sampling it and appends it to the current set of centers. The distribution is updated after each cluster center is selected. In contrast, the K-Means|| algorithm updates the distribution much more infrequently, after choosing  $q$  cluster centers in each iteration. This process is repeated for  $s$  iterations, and finally the number of cluster centers obtained is  $sq$ . The chosen centers are re-clustered to obtain the initial set of  $K$  clusters. It is clear that  $s$  must be chosen such that  $sq > K$ . We provide only the initialization of the weighted K-Means|| algorithm that takes the data distribution,  $\{p_l(\mathbf{x}_i)\}_{i=1}^T$ ,

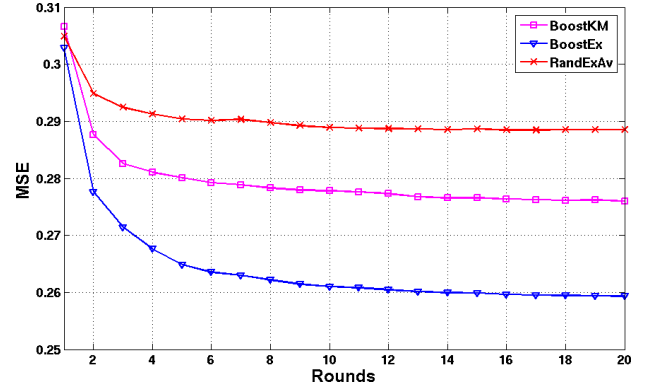


Fig. 3. Convergence characteristics of the proposed ensemble learning approaches.

also into consideration.

Let us denote  $\delta_i$  as the shortest distance of the  $i^{\text{th}}$  training sample to the set of cluster centers already chosen. The initialization of the weighted K-Means|| algorithm proceeds as follows:

- Initialize  $\overline{\mathcal{M}} = \{\}$ .
- Pick the first center  $\mu_1$  from the training set based on the distribution  $\{p_l(\mathbf{x}_i)\}_{i=1}^T$ , and append it to  $\overline{\mathcal{M}}$ .
- The set of intermediate cluster centers,  $\mathcal{M}'$ , is created using  $q$  samples from the data,  $\{\mathbf{x}_i\}_{i=1}^T$ , according to the probability  $\frac{p_l(\mathbf{x}_i)\delta_i^2}{\sum_{j=1}^T p_l(\mathbf{x}_j)\delta_j^2}$ .
- Augment the set  $\overline{\mathcal{M}} \leftarrow \overline{\mathcal{M}} \cup \mathcal{M}'$ .
- Repeat steps 2, 3 and 4 for  $s$  iterations.
- Set the weight of each element  $\mu$  in the set  $\overline{\mathcal{M}}$ , as the sum of weights of samples in  $\mathbf{X}$  that are closer to  $\mu$  than any other sample in  $\overline{\mathcal{M}}$ .
- Perform weighted clustering on the elements of  $\overline{\mathcal{M}}$  to obtain the set of  $K$  cluster centers,  $\mathcal{M}$ .

Note that the steps (b) and (c) are used to compute the initial cluster centers giving preference to samples with higher probability mass. Finally, each dictionary atom  $\mathbf{d}_k$  is set as the normalized cluster center  $\frac{\mu_k}{\|\mu_k\|_2}$ .

2) *BoostEx*: From (3), it is clear that the learned dictionary atoms are close to training samples that have higher probabilities. Therefore, in the *BoostEx* method, the dictionary for round  $l$  is updated by choosing  $K$  data samples based on the non-uniform weight distribution,  $\{p_l(\mathbf{x}_i)\}_{i=1}^T$ , and normalizing them. This scheme will ensure that those samples with high approximation errors in the previous round, will be better represented in the current round.

### C. Demonstration of the Proposed Approaches

The performance of the proposed ensemble schemes for dictionaries of three different sizes  $K = \{256, 1024, 2048\}$  are compared. The training set described in Section II-B is used with the *RandExAv*, *BoostKM*, and *BoostEx* schemes. The dictionaries  $\{\mathbf{D}_l\}_{l=1}^L$  and the weights  $\{\beta_l\}_{l=1}^L$  are obtained with the above schemes for  $L = 20$ . The individual approximations in the training set are obtained using (1) with the sparsity penalty set as  $\lambda = 0.2$ . For each sample in the test set described in Section II-B, the individual representations are

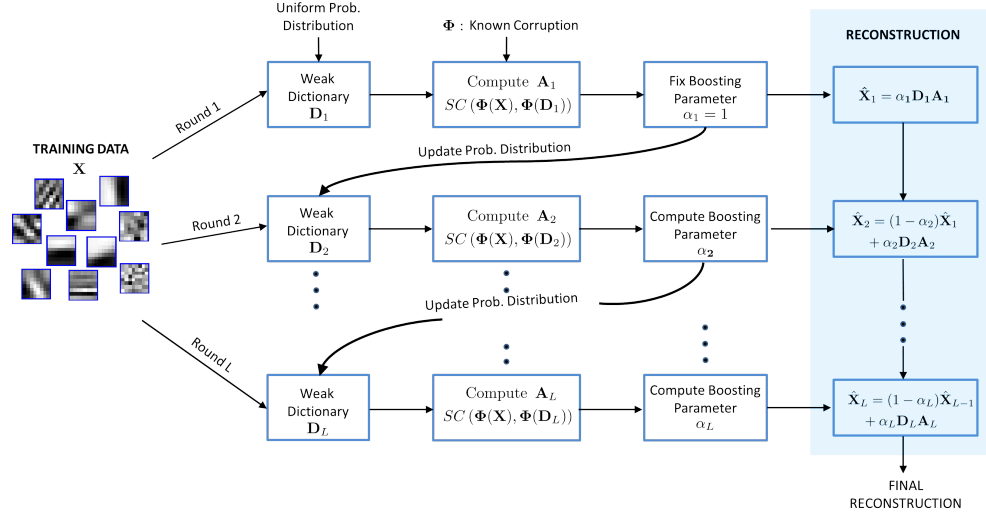


Fig. 4. Illustration of the proposed boosted dictionary learning for image restoration. SC denotes sparse coding using (15).

computed using (1) with  $\lambda = 0.2$ . The ensemble approximation the  $i^{\text{th}}$  test sample is obtained as  $\sum_{l=1}^L \beta_l \mathbf{D}_l \mathbf{a}_{l,i}$ . Figure 2 compares the performances of the proposed schemes for different dictionary sizes. The minimum error obtained across all individual approximations is also shown for comparison, with all the three methods and the different dictionary sizes. It can be seen that the proposed schemes satisfy the basic property of the ensemble discussed in Section II, where it has been shown that the ensemble approximation performs better than the best constituent individual approximation. As the number of number of approximations in the ensemble increase, the average mean squared error (MSE) for the three proposed methods reduce, as shown in Figure 3 for a dictionary size of 1024. Clearly, increasing the number of models in the ensemble results in a better approximation, but the MSE flattens out as the number of rounds increase.

#### IV. APPLICATION: IMAGE RESTORATION

In restoration applications, it is necessary to solve an inverse problem, in order to estimate the test data  $\mathbf{y}$  from

$$\mathbf{z} = \Phi(\mathbf{y}) + \mathbf{n}, \quad (14)$$

where  $\Phi(\cdot)$  is the corruption operator and  $\mathbf{n}$  is the additive noise. If the operator  $\Phi(\cdot)$  is linear, we can represent it using the matrix  $\Phi$ . With the prior knowledge that  $\mathbf{y}$  is sparsely representable in a dictionary  $\mathbf{D}$  according to (1), (14) can be expressed as  $\mathbf{z} = \Phi \mathbf{D} \mathbf{a} + \mathbf{n}$ . Restoring  $\mathbf{x}$  now reduces to computing the sparse codes  $\mathbf{a}$  by solving

$$\min_{\mathbf{a}} \|\mathbf{z} - \Phi \mathbf{D} \mathbf{a}\|_2^2 + \lambda \|\mathbf{a}\|_1. \quad (15)$$

and finally estimating  $\mathbf{y} = \mathbf{D} \mathbf{a}$  [2]. In the proposed ensemble methods, the final estimate of  $\mathbf{x}$  is obtained as a weighted average of the individual approximations. Furthermore, in the boosting approaches, *BoostKM* and *BoostEx*, the degradation operation can be included when learning the ensemble. This is achieved by degrading the training data as  $\Phi \mathbf{X}$ , and obtaining the approximation with the coefficients computed using (15)

instead of (1). The procedure to obtain boosted dictionaries using degraded data and computing the final approximation is illustrated in Figure 4. In this figure, the final approximation is estimated sequentially using the weights  $\{\alpha_l\}_{l=1}^L$ , but it is equivalent to computing  $\{\beta_l\}_{l=1}^L$  using (10) and computing the ensemble estimate  $\sum_{l=1}^L \beta_l \mathbf{D}_l \mathbf{A}_l$ .

##### A. Compressive Recovery

In compressed sensing (CS), the  $N$ -dimensional observation  $\mathbf{z}$  is obtained by projecting the  $M$ -dimensional data  $\mathbf{y}$  onto a random linear subspace, where  $N \ll M$  [42]. In this case, the entries of the degradation matrix  $\Phi \in \mathbb{R}^{N \times M}$  are obtained as i.i.d. realizations of a Gaussian or Bernoulli random variable. Compressive recovery can be effectively performed using conventional dictionaries or ensemble dictionaries. In addition, the proposed idea of ensemble learning can be incorporated in existing learning schemes to achieve improved recovery performance. In particular, the multilevel dictionary learning algorithm [3] can be very easily adapted to compute ensemble representations. Before discussing the experimental setup, and the results of the proposed methods, we will describe the modification to multilevel dictionary learning for improving the compressed recovery performance with learned dictionaries.

1) *Improved Multilevel Dictionaries*: The multilevel dictionary (MLD) learning algorithm is a hierarchical procedure where the dictionary atoms in each level are obtained using a 1-D subspace clustering procedure [3]. Multilevel dictionaries have been shown to generalize well to novel test data, and have resulted in high performance in compressive recovery. We propose to employ the *RandExAv* procedure in each level of multilevel learning to reduce overfitting and thereby improve the accuracy of the dictionaries in representing novel test samples. In each level,  $L$  different dictionaries are drawn as random subsets of normalized training samples. For each training sample, a 1-sparse representation is computed with each individual dictionary, and the approximations are averaged to obtain the ensemble representation for that level.



TABLE I

COMPRESSED RECOVERY OF STANDARD IMAGES: PSNR (dB) OBTAINED USING ALTERNATING DICTIONARY OPTIMIZATION (*Alt-Opt*), BOOSTEX (*BEx*), BOOSTKM (*BKM*), RANDEXAV (*RExAv*), AND *Ex-MLD* METHODS, FOR DIFFERENT VALUES OF  $N$ . THE RESULTS REPORTED WERE OBTAINED BY AVERAGING OVER 10 ITERATIONS WITH DIFFERENT RANDOM MEASUREMENT MATRICES. IN EACH, THE HIGHER PSNR IS GIVEN IN BOLD FONT.

Image	Number of Measurements (N)														
	N = 8					N = 16					N = 32				
	Alt-Opt	BEx	BKM	RExAv	Ex-MLD	Alt-Opt	BEx	BKM	RExAv	Ex-MLD	Alt-Opt	BEx	BKM	RExAv	Ex-MLD
Barbara	21.55	22.05	22.04	22.08	<b>22.95</b>	23.52	23.86	23.73	23.68	<b>24.39</b>	26.45	26.53	26.44	26.28	<b>26.66</b>
Boat	23.08	23.73	23.95	23.99	<b>25.08</b>	25.91	26.29	26.57	26.59	<b>26.96</b>	28.79	29.32	29.61	29.61	<b>29.9</b>
Couple	23.15	23.81	24.02	24.05	<b>25</b>	25.87	26.30	26.57	26.56	<b>27.19</b>	28.83	29.33	29.68	29.67	<b>29.8</b>
Fingerprint	18.10	18.76	19.15	19.16	<b>20.39</b>	21.74	22.18	22.84	22.86	<b>23.19</b>	25.36	25.85	26.59	26.63	<b>26.82</b>
House	24.52	25.12	25.51	25.52	<b>26.55</b>	28.01	28.14	28.63	28.66	<b>28.93</b>	31.28	31.53	32.01	32.03	<b>32.25</b>
Lena	25.14	25.84	26.18	26.25	<b>27.17</b>	28.31	28.73	29.08	29.13	<b>29.59</b>	31.12	31.77	32.15	32.17	<b>32.65</b>
Man	23.90	24.60	24.83	24.89	<b>25.84</b>	26.60	27.12	27.35	27.40	<b>27.68</b>	29.45	30.14	30.40	30.42	<b>30.67</b>
Peppers	21.31	21.83	22.17	22.23	<b>23.12</b>	24.30	24.54	24.82	24.91	<b>25.68</b>	27.28	27.69	28.03	28.11	<b>28.57</b>

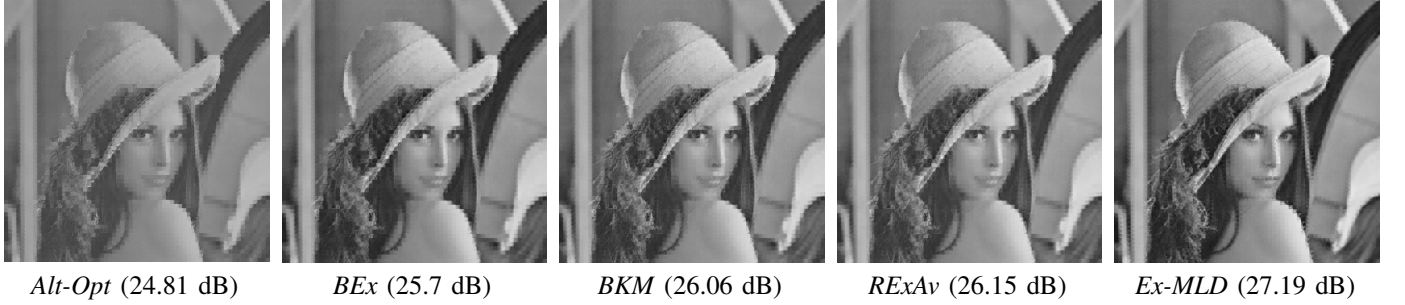


Fig. 5. Compressed recovery of *Man* image using *BoostKM* dictionaries. The reconstructed images along with their corresponding PSNR are shown for the rounds  $\{1, 5, 20, 50\}$ , when 25% random measurements are used.

Using the residual vectors as the training data, this process is repeated for multiple levels. The sparse approximation for a test sample is computed in a similar fashion. Since the sparse code computation in each level is performed using simple correlation operations, the computation complexity is not increased significantly by employing ensemble learning. In our simulations, we will refer to this approach as Example-based Multilevel Dictionary learning (*Ex-MLD*).

2) *Results*: The training set is the same as that described in Section II-B. For the baseline *Alt-Opt* approach, we train a single dictionary with  $K = 256$  using 100 iterations with the sparsity penalty  $\lambda_{tr}$  set to 0.1. The ensemble learning procedures *BoostEx*, *BoostKM* and *RandExAv* are trained with  $L = 50$  and  $K = 256$  for sparsity penalty  $\lambda_{tr} = 0.1$ . The boosted ensembles are trained by taking the random projection operator into consideration, as discussed in Section IV for the reduced measurements,  $N = \{8, 16, 32\}$ . For the *Ex-MLD* method, both the number of levels and the number of atoms in each level were fixed at 16. In each level, we obtained  $L = 50$  dictionaries to compute the ensemble representation.

The recovery performance of the proposed ensemble models is evaluated using the set of standard images shown in Table I. Each image is divided into non-overlapping patches of size  $8 \times 8$ , and random projection is performed with the number of measurements set at  $N = \{8, 16, 32\}$ . For the *Alt-Opt* procedure, the individual patches are recovered using (15), and for the ensemble methods, the approximations computed using the  $L$  individual dictionaries are combined. The penalty  $\lambda_{te}$  is set to 0.1 for sparse coding in all cases. For each method, the PSNR values were obtained by averaging the results over 10

iterations with different random measurement matrices, and the results are reported in Table I. It was observed that the proposed ensemble methods outperform the *Alt-Opt* methods in all cases. In particular, we note that the simple *RandExAv* performs better than the boosting approaches, although in Section II-B it was shown that boosting approaches show a superior performance. The reason for this discrepancy is that boosting aggressively reduces error with training data, and hence may lead to overfitting with degraded test data. Whereas, the *RandExAv* method provides the same importance to all individual approximations both during the training and the testing phases. As a result, it provides a better generalization in the presence of degradation. We also note that similar behavior has been observed with ensemble classification methods [43]. Random sampling methods such as bagging perform better than boosting with noisy examples, since bagging exploits classification noise to produce more diverse classifiers. Furthermore, we observed that the proposed *Ex-MLD* method performed significantly better than all approaches, particularly for lower number of measurements. Figure 5 shows the images recovered using the different approaches, when  $N$  was fixed at 8. As it can be observed, the *Ex-MLD* and *RandExAv* methods provide PSNR gains of 2.38dB and 1.34dB respectively, when compared to the *Alt-Opt* approach.

### B. Single Image Superresolution

Single image superresolution (SISR) attempts to reconstruct a high-resolution image using just a single low-resolution image. It is a severely ill-posed problem and in sparse representation based approaches, the prior knowledge that

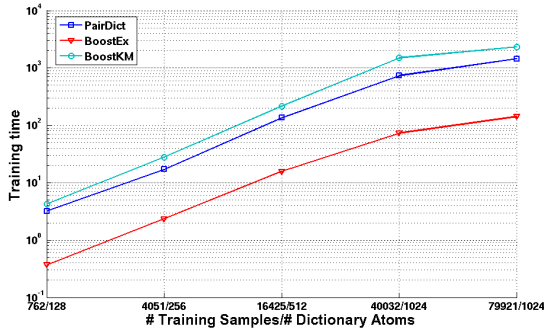


Fig. 6. Effect of dictionary and training set sizes on the dictionary training time for different learning schemes. The training times given are in seconds and are compared only for *PairDict* (40 iterations), *BoostEx* ( $L = 50$ ) and *BoostKM* ( $L = 50$ ) since *ExDict* and *RandExAv* require no training.

natural image patches can be represented as a sparse linear combination of elementary patches, is used. The degraded test image is represented as  $\mathbf{Z} = \Phi \mathbf{Y}$ , where the operator  $\Phi$  blurs the high-resolution image  $\mathbf{Y}$  and then downsamples it. Note that  $\mathbf{Y}$  and  $\mathbf{Z}$  denote vectorized high- and low-resolution images respectively. Each overlapping patch obtained from the degraded image is denoted as  $\mathbf{z}$ . The paired dictionary learning procedure (*PairDict*) proposed in [36] has been very effective in recovering the high-resolution patches. This method initially creates degraded counterparts of the high-resolution training images, following which gradient-based features are extracted from the low-resolution patches and the features are appended to the corresponding vectorized high-resolution patches. These augmented features are used to train a paired dictionary  $(\mathbf{D}_{lo}, \mathbf{D}_{hi})$  such that each low-resolution and its corresponding high-resolution patches share the same sparse code. For a low-resolution test patch  $\mathbf{z}$ , the sparse code  $\mathbf{a}$  is obtained using  $\mathbf{D}_{lo}$ , and the corresponding high-resolution counterpart is recovered as  $\mathbf{y} = \mathbf{D}_{hi} \mathbf{a}$ . An initial estimate  $\mathbf{Y}_0$  of the high-resolution image is obtained by appropriately averaging the overlapped high-resolution patches. Finally, a global reconstruction constraint is enforced by projecting  $\mathbf{Y}_0$  on to the solution space of  $\Phi \mathbf{Y} = \mathbf{Z}$ ,

$$\min_{\mathbf{Y}} \|\mathbf{Z} - \Phi \mathbf{Y}\|_2^2 + c \|\mathbf{Y} - \mathbf{Y}_0\|_2^2, \quad (16)$$

to obtain the final reconstruction. As an alternative, the example-based procedure (*ExDict*) proposed in [20], the dictionaries  $\mathbf{D}_{lo}$  and  $\mathbf{D}_{hi}$  are directly fixed as the features extracted from low-resolution patches and vectorized high resolution patches respectively. Similar to the *PairDict* method, the global reconstruction constraint in (16) is imposed for the final reconstruction.

In our simulations, standard grayscale images (Table II) are magnified by a factor of 2, using the proposed approaches. In addition to the *PairDict* and *ExDict* methods, simple bicubic interpolation is also used as a baseline method. We also obtained paired dictionaries with 1024 atoms using 100,000 randomly chosen patches of size  $5 \times 5$  from the grayscale natural images in the training set. The sparsity penalty used in training was  $\lambda_{tr} = 0.15$ . The training set was reduced to the size of 20,000 samples and used as the dictionary for the

TABLE II  
SUPERRESOLUTION OF STANDARD IMAGES UPSCALED BY A FACTOR OF 2: PSNR IN dB OBTAINED WITH BICUBIC INTERPOLATION (*Bicubic*), PAIRED DICTIONARY (*PairDict*) [36], EXAMPLE DICTIONARY (*ExDict*) [20], BOOSTEX (*BEx*), BOOSTKM (*BKM*), AND RANDEXAV (*RExAv*) METHODS.

Image	Bicubic	PairDict	ExDict	BEx	BKM	RExAv
Lena	34.10	<b>35.99</b>	<b>35.99</b>	35.97	35.95	<b>35.99</b>
Boat	29.94	<b>31.34</b>	<b>31.34</b>	31.28	31.23	31.29
House	32.77	<b>34.49</b>	<b>34.49</b>	34.38	34.41	34.41
Cameraman	26.33	<b>27.78</b>	<b>27.78</b>	27.72	<b>27.78</b>	27.71
Straw	24.20	25.93	25.93	25.90	25.90	<b>25.94</b>
Girl	33.81	<b>35.39</b>	<b>35.39</b>	35.33	35.37	35.35

*ExDict* method. For ensemble learning,  $L$  was fixed at 50 and the approximation for each data sample was obtained using just a 1-sparse representation.

For different number of training samples, we compared the training times for *PairDict* (40 iterations), *BoostEx* ( $L = 50$ ) and *BoostKM* ( $L = 50$ ) algorithms in Figure 6. The computation times reported in this paper were obtained using a single core of a 2.8 GHz Intel i7 Linux machine with 8GB RAM. The *BoostKM* approach has the maximum computational complexity for training, followed by *PairDict* and *BoostEx* approaches. The *ExDict* procedure requires no training and for *RandExAv*, training time is just the time for randomly selecting  $K$  samples from the training set of  $T$  samples, for  $L$  rounds. Clearly, the complexity incurred for this is extremely low.

For the test images, SISR is performed using the baseline *PairDict* and *ExDict* approaches using a sparsity penalty of  $\lambda_{te} = 0.2$ . For the *PairDict*, and *ExDict* approaches, the code provided by the authors [38] was used to generate the results. The recovery performance of the proposed algorithms are reported in Table II. For *PairDict*, as well the proposed ensemble methods, the dictionary size is fixed at 1024, whereas all the examples are used for training with the *ExDict* approach. We observed from our results that an ensemble representation with a simplified sparse coding scheme (1-sparse) matched the performance of the baseline methods (Figure 7).

## V. APPLICATION: UNSUPERVISED CLUSTERING

Conventional clustering algorithms such as K-Means provide good clusterings only when the natural clusters of the data are distributed around a mean vector in space. For data that lie in a union of low-dimensional subspaces, it is beneficial to develop clustering algorithms that try to model the actual data distribution better. The sparse subspace clustering method [44], a special case of which is referred to as the  $\ell_1$  graph clustering [29], results in clusters that correspond to subspaces of data. This is achieved by representing each example as a sparse linear combination of the others and finally performing spectral clustering using a similarity matrix obtained from the coefficient matrix. The clustering method has the advantage of incorporating the noise model directly when performing sparse coding, thereby achieving robustness. The coefficient vector for the  $i^{\text{th}}$  data sample is obtained as

$$\min_{\mathbf{b}_i} \|\mathbf{x}_i - \mathbf{X} \mathbf{a}_i\| + \lambda \|\mathbf{a}_i\|_1, \text{ subj. to. } a_{ii} = 0. \quad (17)$$





Fig. 7. SISR of the *Man* image with scaling factor of 2. The *PairDict*, *ExDict*, and *RandExAv* methods result in very similar high resolution images.

By imposing the constraint that the  $i^{\text{th}}$  element of  $\mathbf{a}_i$  should be 0, we ensure that a data sample is not represented by itself, which would have resulted in a trivial approximation. The coefficient matrix is denoted as  $\mathbf{A} = [\mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_T]$ , and spectral clustering [45] is performed by setting the similarity matrix to the symmetric non-negative version of the coefficient matrix,  $\mathbf{S} = |\mathbf{A}| + |\mathbf{A}^T|$ . Computing the graph in this case necessitates the computation of sparse codes of  $T$  data samples with a  $M \times (T - 1)$  dictionary. Sparse coding-based graphs can also be constructed based on coefficients obtained with a dictionary  $\mathbf{D}$ , inferred using the *Alt-Opt* procedure. Denoting the sparse codes for the examples  $\mathbf{X}$  by the coefficient matrix  $\mathbf{A} = [\mathbf{a}_1 \dots \mathbf{a}_T]$ , the similarity matrix can be constructed as  $\mathbf{S} = |\mathbf{A}^T \mathbf{A}|$ . Similar to the  $\ell_1$  graphs, this similarity matrix can be used with spectral clustering to estimate the cluster memberships [5]. In this case, the dominant complexity in computing the graph is in learning the dictionary, and obtaining the sparse codes for each example. When the number of training examples is large, or when the data is high-dimensional, approaches that use sparse coding-based graphs incur high computational complexity.

We propose to construct sparse representation-based graphs using our ensemble approaches and employ them in spectral clustering. In our ensemble approaches, we have two example-based procedures, (*RandExAv* and *BoostEx*) and one that uses K-Means dictionaries (*BoostKM*). For *BoostKM*, we obtain  $L$  dictionaries of size  $K$  using the boosting procedure, with 1-sparse approximations. The final coefficient vector of length  $LK$  for the data sample  $\mathbf{x}_i$  is obtained as,  $\mathbf{a}_i = [\mathbf{a}_{1,i}^T \mathbf{a}_{2,i}^T \dots \mathbf{a}_{L,i}^T]^T$ , where  $\mathbf{a}_{l,i}$  is the coefficient vector for round  $l$ . The similarity matrix is then estimated as  $\mathbf{S} = |\mathbf{A}^T \mathbf{A}|$ . In the example-based procedures, again 1-sparse representation is used to obtain the coefficient vectors  $\{\mathbf{a}_{1,i}, \mathbf{a}_{2,i} \dots, \mathbf{a}_{L,i}\}$ , for a data sample  $\mathbf{x}_i$ . A cumulative coefficient vector of length  $T$  can be obtained by recognizing that each coefficient in  $\mathbf{a}_{l,i} \in \mathbb{R}^K$ , can be associated to a particular example, since  $\mathbf{D}_l$  is an example-based dictionary. Therefore a new 1-sparse coefficient vector  $\bar{\mathbf{a}}_{l,i} \in \mathbb{R}^T$  is created such that  $\mathbf{D}_l \mathbf{a}_{l,i} = \bar{\mathbf{X}} \bar{\mathbf{a}}_{l,i}$ , where  $\bar{\mathbf{X}}$  contains the normalized set of data samples  $\mathbf{X}$ . Finally the cumulative coefficient vector for  $\mathbf{x}_i$  is obtained as  $\sum_{l=1}^L \beta_l \bar{\mathbf{a}}_{l,i}$ . They are then stacked to form the coefficient matrix  $\bar{\mathbf{A}} = [\bar{\mathbf{a}}_1 \dots \bar{\mathbf{a}}_T]$ . Spectral clustering can be now performed using the similarity matrix  $\mathbf{S} = |\bar{\mathbf{A}}| + |\bar{\mathbf{A}}^T|$ . The clustering performance was evaluated in terms of accuracy and normalized mutual information (NMI), and compared with  $\ell_1$  graphs. As seen

TABLE III  
COMPARISON OF THE CLUSTERING PERFORMANCES (ACCURACY AND NORMALIZED MUTUAL INFORMATION) OF THE ALGORITHMS WITH STANDARD DATASETS. THE BEST MAXIMUM OR AVERAGE PERFORMANCE IS GIVEN IN BOLD FONT.

Dataset	$\ell_1$ graph		BoostEx		RandExAv		BoostKM	
	max	avg	max	avg	max	avg	max	avg
<b>Accuracy</b>								
Digits	<b>88.72</b>	74.61	88.63	76.79	88.62	<b>77.56</b>	88.31	76.85
Soybean	67.44	58.22	67.26	63.21	<b>70.82</b>	<b>65.31</b>	69.22	63.50
Segment	65.32	57.84	63.42	58.46	63.33	56.49	<b>65.63</b>	<b>58.67</b>
Satimage	77.56	69.37	75.03	72.88	<b>83.59</b>	<b>75.25</b>	71.62	65.32
USPS	78.24	62.47	75.01	68.18	78.26	64.14	<b>90.96</b>	<b>75.18</b>
<b>NMI</b>								
Digits	<b>84.97</b>	76.50	84.69	77.67	84.73	<b>78.04</b>	84.31	78.02
Soybean	74.55	65.94	72.84	68.66	<b>77.50</b>	<b>73.71</b>	76.05	71.66
Segment	<b>59.14</b>	53.91	56.72	53.23	54.94	51.20	58.44	<b>55.32</b>
Satimage	65.09	61.20	62.98	59.01	<b>69.38</b>	<b>66.12</b>	60.27	55.57
USPS	81.04	74.14	70.52	66.89	82.67	76.95	<b>82.78</b>	<b>78.37</b>

from Table III, the ensemble-based approaches result in high accuracy as well as NMI. In all our simulations, data was preprocessed by centering and normalizing to unit norm. It was observed that the proposed ensemble methods incur comparable computational complexity to  $\ell_1$  graphs for datasets with small data dimensions. However, we observed significant complexity reduction with the *USPS* dataset, which contains 9298 samples of 256 dimensions. To cluster the *USPS* samples, the  $\ell_1$  graph approach took 411.85 seconds to compute the sparse codes, whereas *BoostEx*, *RandExAv*, and *BoostKM* took 152.56, 147.93, and 83.58 seconds respectively. This indicates the suitability of the proposed methods for high-dimensional, large scale data.

## VI. CONCLUSIONS

We proposed and analyzed the framework of ensemble sparse models, where the data is represented using a linear combination of approximations from multiple sparse representations. Theoretical results and experimental demonstrations show that an ensemble representation leads to a better approximation when compared to its individual constituents. Three different methods for learning the ensemble were proposed. Results in compressive recovery showed that the proposed approaches performed better than the baseline sparse coding method. Furthermore, the ensemble approach performed comparably to several recent techniques in single image superresolution. Results with unsupervised clustering also showed that

the proposed method leads to better clustering performance in comparison to the  $\ell_1$  graph method.

## REFERENCES

- [1] D. Field, "What is the goal of sensory coding?" *Neural computation*, vol. 6, no. 4, pp. 559–601, 1994.
- [2] M. Elad, *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer, 2010.
- [3] J. Thiagarajan, K. Ramamurthy, and A. Spanias, "Multilevel dictionary learning for sparse representation of images," in *IEEE DSPE Workshop*, 2011, pp. 271–276.
- [4] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.
- [5] I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in *IEEE CVPR*, 2010, pp. 3501–3508.
- [6] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Supervised dictionary learning," *Advances in neural information processing systems*, 2008.
- [7] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via  $\ell_1$  minimization," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 5, pp. 2197–202, Mar. 2003.
- [8] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.
- [9] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2231–2242, October 2004.
- [10] S. Cotter, R. Adler, R. Rao, and K. Kreutz-Delgado, "Forward sequential algorithms for best basis selection," in *Vision, Image and Signal Processing, IEE Proceedings-*, vol. 146, no. 5. IET, 1999, pp. 235–244.
- [11] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Review*, vol. 43, no. 1, pp. 129–159, 2001.
- [12] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using FOCUSS: A re-weighted norm minimization algorithm," *IEEE Transactions on Signal Processing*, vol. 45, no. 3, pp. 600–616, March 1997.
- [13] M. Elad, "Why simple shrinkage is still relevant for redundant representations?" *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5559–5569, December 2006.
- [14] M. Elad *et al.*, "A wide-angle view at iterated shrinkage algorithms," in *SPIE (Wavelet XII) 2007*, 2007.
- [15] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proc. ICML*, 2009, pp. 689–696.
- [16] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of statistics*, vol. 32, no. 2, pp. 407–499, 2004.
- [17] H. Lee, A. Battle, R. Raina, and A. Ng, "Efficient sparse coding algorithms," *Advances in neural information processing systems*, vol. 19, p. 801, 2007.
- [18] M. Aharon, M. Elad, and A. Bruckstein, "The K-SVD: an algorithm for designing of overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.
- [19] M. Aharon and M. Elad, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3736–3745, 2006.
- [20] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution as sparse representation of raw image patches," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.
- [21] K. Huang and S. Aviyente, "Sparse representation for signal classification," in *Proc. of Advances in Neural Information Processing Systems 19*. MIT Press, 2006.
- [22] J. J. Thiagarajan, K. N. Ramamurthy, and A. Spanias, "Sparse representations for pattern classification using learned dictionaries," *Proc. of Twenty-eighth SGAI International Conference on Artificial Intelligence*, 2008.
- [23] J. J. Thiagarajan, K. N. Ramamurthy, P. Knee, and A. Spanias, "Sparse representations for automatic target classification in SAR images," in *Proc. of ISCCSP*, 2010.
- [24] J. Yang *et al.*, "Linear spatial pyramid matching using sparse coding for image classification," in *IEEE CVPR*, 2009.
- [25] G. Yu, G. Sapiro, and S. Mallat, "Image modeling and enhancement via structured sparse model selection," in *Proc. of IEEE ICIP*, Sep. 2010, pp. 1641–1644.
- [26] Q. Zhang and B. Li, "Discriminative K-SVD for dictionary learning in face recognition," in *IEEE CVPR*, 2010.
- [27] J. J. Thiagarajan and A. Spanias, "Learning dictionaries for local sparse coding in image classification," in *Proc. of Asilomar SSC*, 2011.
- [28] J. J. Thiagarajan, K. N. Ramamurthy, P. Sattigeri, and A. Spanias, "Supervised local sparse coding of sub-image features for image retrieval," in *IEEE ICIP*, 2012.
- [29] B. Cheng, J. Yang, S. Yan, Y. Fu, and T. S. Huang, "Learning with  $\ell_1$ -graph for image analysis," *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, vol. 19, no. 4, pp. 858–66, Apr. 2010. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/20031500>
- [30] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *Journal-Japanese Society For Artificial Intelligence*, vol. 14, no. 771–780, p. 1612, 1999.
- [31] R. Polikar, "Ensemble based systems in decision making," *Circuits and Systems Magazine, IEEE*, vol. 6, no. 3, pp. 21–45, 2006.
- [32] W. Zhang, A. Surve, X. Fern, and T. Dietterich, "Learning non-redundant codebooks for classifying complex objects," in *Proc. ICML*, 2009, pp. 1241–1248.
- [33] J. Wang, Y. Li, Y. Zhang, H. Xie, and C. Wang, "Boosted learning of visual word weighting factors for bag-of-features based medical image retrieval," in *International Conference on Image and Graphics*, 2011, pp. 1035–1040.
- [34] M. Elad and I. Yavneh, "A plurality of sparse representations is better than the sparsest one alone," *Information Theory, IEEE Transactions on*, vol. 55, no. 10, pp. 4701–4714, 2009.
- [35] N. Duffy and D. Helmbold, "Boosting Methods for Regression," pp. 153–200, 2002.
- [36] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *Image Processing, IEEE Transactions on*, vol. 19, no. 11, pp. 2861–2873, 2010.
- [37] T. Dietterich, "Ensemble methods in machine learning," *Multiple classifier systems*, pp. 1–15, 2000.
- [38] "ScSR - matlab codes for image super-resolution," Available at <http://www.ifp.illinois.edu/~jyang29/resources.html>.
- [39] "Berkeley segmentation dataset," Available at <http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/segbench/>.
- [40] B. Bahmani, B. Moseley, A. Vattani, R. Kumar, and S. Vassilvitskii, "Scalable k-means++," *Proceedings of the VLDB Endowment*, vol. 5, no. 7, pp. 622–633, 2012.
- [41] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proc. ACM-SIAM symposium on Discrete algorithms*, 2007, pp. 1027–1035.
- [42] D. Donoho, "Compressed sensing," *IEEE Trans. Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.
- [43] T. G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization," *Machine learning*, vol. 40, no. 2, pp. 139–157, 2000.
- [44] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *arXiv preprint arXiv:1203.1005*, 2012.
- [45] A. Y. Ng, M. I. Jordan, Y. Weiss *et al.*, "On spectral clustering: Analysis and an algorithm," *Advances in neural information processing systems*, vol. 2, pp. 849–856, 2002.